Neuropsychological deficits in frontotemporal dementia and Alzheimer’s disease: a meta-analytic review

A D Hutchinson and J L Mathias

J. Neurol. Neurosurg. Psychiatry 2007;78:917-928; originally published online 19 Mar 2007;
doi:10.1136/jnnp.2006.100669

Updated information and services can be found at:
http://jnnp.bmj.com/cgi/content/full/78/9/917

These include:

References
This article cites 131 articles, 44 of which can be accessed free at:
http://jnnp.bmj.com/cgi/content/full/78/9/917#BIBL

2 online articles that cite this article can be accessed at:
http://jnnp.bmj.com/cgi/content/full/78/9/917#otherarticles

Rapid responses
You can respond to this article at:
http://jnnp.bmj.com/cgi/eletter-submit/78/9/917

Email alerting service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article

Notes

To order reprints of this article go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to Journal of Neurology, Neurosurgery, and Psychiatry go to:
http://journals.bmj.com/subscriptions/
Neuropsychological deficits in frontotemporal dementia and Alzheimer’s disease: a meta-analytic review

A D Hutchinson, J L Mathias

We sought to identify the cognitive tests that best discriminate between Alzheimer’s disease (AD) and frontotemporal dementia (FTD). A comprehensive search of all studies examining the cognitive performance of persons diagnosed with AD and FTD, published between 1980 and 2006, was conducted. Ninety-four studies were identified, comprising 2936 AD participants and 1748 FTD participants. Weighted Cohen’s $d$ effect sizes, percentage overlap statistics, confidence intervals and fail-safe $N$s were calculated for each cognitive test that was used by two or more studies. The most discriminating cognitive tests were measures of orientation, memory, language, visuomotor function and general cognitive ability. Although there were large and significant differences between groups on these measures, there was substantial overlap in the scores of the AD and FTD groups. Age, education, years since diagnosis and diagnostic criteria did not significantly contribute to the group differences. Given the large overlap in the test performance of persons diagnosed with AD and FTD, cognitive tests should be used cautiously and in conjunction with a medical history, behavioural observations, imaging and information from relatives when making differential diagnoses.

Dementia is an important social and health issue in our ageing population, with the prevalence of Alzheimer’s disease (AD) in the USA predicted to increase by approximately 300% by 2050.1 Two forms of dementia, AD and frontotemporal dementia (FTD), and their differential diagnosis using cognitive tests, are addressed in the current study.

AD is the most prevalent and well researched type of dementia.5 People with AD usually exhibit a progressive decline in general cognitive functioning,1 together with acalculia, apraxia, visuospatial problems, poor orientation and impaired language comprehension, attention and memory.2 4 6 Memory problems are one of the earliest symptoms, with the ability to learn and retrieve information being affected initially and impairments in recognition memory developing later.2 6 7

FTD, on the other hand, is variously referred to as Pick disease, dementia of the frontal type, frontal lobe dementia of the non-Alzheimer’s type, frontotemporal lobar degeneration, semantic dementia and progressive non-fluent aphasia.5 6 FTD refers to a group of degenerative dementias that are characterised by atrophy of the frontotemporal cortex.5 6 10 Although people with FTD may also experience memory problems, these deficits are more frequently associated with a tendency to respond impulsively and a failure to monitor performance.5 9 As with AD, retrieval of information from memory is affected. However, unlike AD, the provision of cues can assist performance.7 Persons with FTD also have more impairments in executive functioning5 6 9 11 12 and language (eg, inability to repeat, stereotyped phrases, reduced speech).5 7 8 Finally, although general cognitive ability may be unaffected in the early stages of FTD, it does decline over time.4 12

While definitive diagnoses of AD and FTD can only be made at autopsy, interim diagnoses usually take into account the base rates of the disorder, clinical criteria, medical history, physical examination, brain imaging and neuropsychological assessments.4 10 11 Unfortunately, the diagnosis of FTD can be difficult because of its insidious and gradual onset5 9 and can also be misdiagnosed as AD.13 14 However, accurate differential diagnosis has become increasingly important because of the recent availability of pharmacological treatments that temporarily improve the cognitive and functional abilities of people with AD.15–21 Moreover, delays in the commencement of these treatments significantly reduce their benefits.19 22 Highlighting the importance of early and accurate diagnosis. In contrast, no specific pharmacological treatments currently exist for FTD, although research is continuing.23

If neuropsychological assessments are to make a valuable contribution to the diagnosis of dementia, clinicians need to know what cognitive tests best discriminate between the different types of dementia. Despite the differing clinical characteristics of these two disorders and the wide variety of tests that have been researched with AD and FTD samples, research has yielded inconsistent findings, both in terms of the cognitive constructs (eg, memory, language) and the specific tests that appear to differentiate between AD and FTD.14 24–33 These inconsistencies may, in part, be due to methodological differences in the research literature, such as variations in the participants’ age, time since diagnosis, stage of dementia and diagnostic criteria. A recent narrative review of the literature concluded that there are no tests that

Abbreviations: ACE, Addenbrooke’s Cognitive Examination; AD, Alzheimer’s disease; AVLT, Auditory Verbal Learning Test; FTD, frontotemporal dementia; MMSE, Mini-Mental State Examination; Ns, fail safe N; %OL, percentage overlap; WAB, Western Aphasia Battery; WLS, weighted least squares; WMS, Wechsler Memory Scale

Received 23 June 2006
Revised 13 February 2007
Accepted 1 March 2007
Published Online First 19 March 2007
accurately discriminate between AD and FTD. However, a review of this type cannot adequately consolidate and analyse the data or reconcile differences in methodology. A meta-analysis, in contrast, provides an objective and quantitative means by which to directly compare existing findings, thereby providing an important addition to the existing literature. The current study therefore undertook a meta-analytic review of research that has compared the cognitive performance of AD and FTD samples in order to identify the tests that best discriminate between AD and FTD and, therefore, are most likely to be of use when making a differential diagnosis.

METHOD

Literature search and inclusion criteria

A comprehensive search of the PubMed and PsycINFO electronic databases from January 1980 to November 2006 was undertaken to identify all published studies that assessed the cognitive functioning of AD and FTD samples. The key search terms that were used to capture relevant articles are shown in table 1. The bibliographies of relevant papers were also examined for additional references. To be selected for the current meta-analysis, a study had to meet the following inclusion criteria: (1) it examined groups of participants with AD and FTD (excludes case studies), (2) cognitive tests were administered to both groups (excludes rating scales and questionnaires), (3) the cognitive tests were not used for the diagnosis and classification of participants into the AD and FTD groups (ie, a test could not be used as both a dependent and independent variable), (4) statistical data enabling the calculation of Cohen’s d effect sizes was provided (eg, mean and SD, t tests or one way F tests) and (5) it was published in English.

This literature search yielded 2785 potentially relevant studies, 93 of which met all of the inclusion criteria. These 93 studies used a total of 136 different cognitive tests which, in turn, yielded results for 1019 different test scores (ie, some tests yielded multiple scores). A total of 114 of these test scores were used by more than one study and were, therefore, included in the current meta-analysis. Of the studies that did not meet one or more of the inclusion criteria, 2687 did not undertake cognitive testing with the AD and FTD samples, 67 did not provide data that would enable the calculation of effect sizes and 96 studies only used cognitive tests for the diagnosis and classification of participants. An additional paper was excluded because it presented data that had been reported in another publication that was to be included in this study. This was done to ensure that the data were independent, as meta-analytic techniques assume that a single study only contributes one score to the calculation of a particular effect size. Finally, one publication was treated as two studies because data were provided for separate samples in the one article. Thus 94 independent studies were effectively included in the final meta-analysis.

Data collection and preparation

Consistent with the most commonly used diagnostic criteria for AD, the Mini-Mental State Examination (MMSE) was frequently used when diagnosing dementia. The MMSE was therefore only included as a dependent variable in the current meta-analysis when groups were formed on the basis of a post-mortem diagnosis of AD or FTD dementia. However, where provided, MMSE scores are reported as descriptive data for the AD and FTD samples (refer to table 2). In addition, whenever it was not clear whether a particular cognitive test was used to diagnose dementia and classify participants, this test was excluded from the meta-analysis in order to ensure that a measure was not used as both a dependent and independent variable.

Where test scores were obtained from different editions of the same test (eg, Wechsler Adult Intelligence Scales), these were combined for the purposes of calculating mean effect sizes. Care was taken to ensure that the scores for a given test provided independent measures of performance. Thus subscores and total scores from a test could not both be used in the calculation of an effect size. Similarly, where a study provided multiple scores for the same test (eg, easy and hard versions of a geometric figure task), effect sizes were calculated for each score and then averaged to provide a single effect size for that study; thereby ensuring that a study only contributed one score to the calculation of a mean effect size. This was not necessary when effect sizes for different scores from the same test were reported separately (eg, speed and accuracy).

All tests were broadly grouped into seven cognitive categories in order to synthesise the findings of this meta-analysis. Lezak et al45 and Spreen and Strauss46 were consulted for this purpose. These categories are: orientation and attention, memory, construction, verbal abilities and language, concept formation and reasoning, executive functioning and motor tasks, and general ability and intelligence.

Effect size calculations and analyses

Cohen’s d effect sizes were calculated to measure the extent of the difference between the scores of the AD and FTD samples, where a small effect size is defined as $d = 0.2$, a medium effect as $d = 0.5$ and a large effect as $d = 0.8$. To put this into perspective, an effect size of 0.5 indicates that the scores for the two groups differ by half of a pooled standard deviation. The percentage overlap between the scores of the two groups can also be calculated from an effect size; $d = 0$ equates to 100% overlap (ie, the groups are indistinguishable), $d = 1.0$ equates to 45% overlap and $d = 3.0$ equates to less than 5% overlap in the groups’ scores. These statistics are provided for all effect sizes in order to facilitate their interpretation, especially with regard to their potential for assisting with differential diagnosis.

Effect sizes were calculated in a multi-stage process. The first stage involved calculating effect sizes for each score of every test that was used by each individual study. When calculating effect sizes, FTD scores were always subtracted from AD scores. In most cases, higher test scores indicated better performance. Therefore, a positive effect size indicates that FTD participants were more impaired on a given measure than AD participants. In cases where a higher score indicated greater impairment than a lower score (eg, errors, speed), the direction of the effect sizes for these scores was transformed so that a positive effect size still indicated greater impairment in the FTD group.

<table>
<thead>
<tr>
<th>Table 1 Key search terms</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alzheimer’s disease</strong></td>
</tr>
<tr>
<td>Alzheimer*</td>
</tr>
<tr>
<td>Alzheimer’s disease</td>
</tr>
<tr>
<td>DAT</td>
</tr>
<tr>
<td>AD</td>
</tr>
<tr>
<td>FTLD</td>
</tr>
<tr>
<td>FTD</td>
</tr>
<tr>
<td>DFT</td>
</tr>
</tbody>
</table>

AD, Alzheimer’s disease; DAT, dementia of the Alzheimer’s type; DFT, dementia of the frontal type; FLD, frontal lobe dementia of the non-Alzheimer’s type; FTLD, frontotemporal lobe dementia; PNFA, progressive non-fluent aphasia; PPA, primary progressive aphasia.

*Search term and its derivatives were used (eg, cognition, cognitive, neuropsychology, neuropsychological).
The effect sizes for all studies that used a particular test score were aggregated to calculate a mean effect size (and SD). Whereas mean effect sizes measure the extent (and direction) of the difference between AD and FTD, the SD indicates the variation in effect sizes between studies. Before aggregating the scores from individual studies, each effect size was weighted to take into account the fact that the reliability of an effect size is affected by the size of the sample from which it is derived. According to Lipsey and Wilson, the inverse of the variance provides a better measure of the precision of an effect size than sample size. Consequently, the inverse of the variance was used to weight all effect sizes ($d_w$) using the formulae below.

$$d_w = \frac{\sum w_{ES}}{\sum w}$$

where $ES$ = effect size, $w = \frac{1}{SE^2}$ and $SE$ = standard error.

Effect sizes are only reported for tests that were used by two or more studies because effect sizes that are based on a single study are not thought to provide a reliable measure of group differences. Ninety-five per cent confidence intervals (CIs) for these mean effect sizes were additionally calculated in order to determine their statistical significance. If a CI does not span zero, this indicates that the true population effect size differs significantly from zero, indicating that there is a significant difference between the performance of the AD and FTD groups.

One problem facing all meta-analyses is that studies with statistically significant results are more likely to be published and, therefore, more likely to be included in a meta-analysis. Failure to include unpublished studies with non-significant results increases the risk of a type 1 error, which may result in an effect size being overestimated. A fail safe $N$ (Nfs) was therefore calculated, using the method described by Rosenthal, to address this problem. This statistic estimates the number of unpublished studies, with non-significant results (ie, small effect) that would be required in order to call the current findings into question. The larger the number, the more confident we can be in a particular finding.

An additional problem that faces meta-analytic studies (and qualitative reviews) is that there are methodological differences between studies that may affect the findings and, therefore, the effect sizes. The impact of diagnostic certainty was therefore examined in order to determine whether the research findings were affected by the criteria that were used to establish

Table 2 Demographic details for the Alzheimer’s disease and frontotemporal dementia groups

<table>
<thead>
<tr>
<th>AD</th>
<th>FTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>N* studies</td>
<td>N* participants</td>
</tr>
<tr>
<td>N</td>
<td>94</td>
</tr>
<tr>
<td>Age (y)</td>
<td>88</td>
</tr>
<tr>
<td>Education (y)</td>
<td>61</td>
</tr>
<tr>
<td>Time since diagnosis (y)</td>
<td>38</td>
</tr>
<tr>
<td>MMSE</td>
<td>76</td>
</tr>
<tr>
<td>CDR</td>
<td>20</td>
</tr>
</tbody>
</table>

AD, Alzheimer’s disease; CDR, Clinical Dementia Rating scale; FTD, frontotemporal dementia; MMSE, Mini-Mental State Examination.

*The N differs between variables because not all studies provided data for each variable.

Table 3 Orientation and attention: weighted Cohen’s d effect sizes* for each test, rank ordered by size

<table>
<thead>
<tr>
<th>Test</th>
<th>N studies</th>
<th>N participants</th>
<th>Mean Cohen’s $d_w$</th>
<th>SD $d_w$ (95% CI)</th>
<th>Min d</th>
<th>Max d</th>
<th>Nfs</th>
<th>%OL</th>
<th>Study references</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE Orientation</td>
<td>2</td>
<td>170</td>
<td>-1.08</td>
<td>0.25 (-1.43, -0.73)</td>
<td>-1.20</td>
<td>-0.95</td>
<td>9</td>
<td>41</td>
<td>80 102</td>
</tr>
<tr>
<td>ACE Attention</td>
<td>2</td>
<td>170</td>
<td>-0.79</td>
<td>0.24 (-1.13, -0.45)</td>
<td>-0.86</td>
<td>-0.71</td>
<td>6</td>
<td>53</td>
<td>74 120</td>
</tr>
<tr>
<td>Trails B (errors)</td>
<td>4</td>
<td>62</td>
<td>-0.60</td>
<td>0.37 (-1.11, -0.09)</td>
<td>-0.75</td>
<td>-0.20</td>
<td>4</td>
<td>62</td>
<td>31 40 84 131</td>
</tr>
<tr>
<td>Stroop (word reading/Stroop A)</td>
<td>4</td>
<td>105</td>
<td>0.46</td>
<td>0.45 (0.02, 0.89)</td>
<td>-0.40</td>
<td>5.2</td>
<td>5</td>
<td>67</td>
<td>56 62 84 96 94</td>
</tr>
<tr>
<td>Digit Symbol</td>
<td>5</td>
<td>373</td>
<td>-0.40</td>
<td>0.33 (-0.69, -0.11)</td>
<td>-0.67</td>
<td>-0.28</td>
<td>5</td>
<td>73</td>
<td>23 93 87 98</td>
</tr>
<tr>
<td>Modified Trails (connection/min)</td>
<td>4</td>
<td>225</td>
<td>-0.38</td>
<td>0.28 (-0.66, -0.10)</td>
<td>-0.69</td>
<td>-0.18</td>
<td>4</td>
<td>73</td>
<td>26 44 40 95 97 98</td>
</tr>
<tr>
<td>Digit Span (number correct—incorrect trail)</td>
<td>6</td>
<td>358</td>
<td>-0.29</td>
<td>0.29 (-0.52, -0.06)</td>
<td>-0.57</td>
<td>0.41</td>
<td>3</td>
<td>79</td>
<td>69 102</td>
</tr>
<tr>
<td>Digit Span (overall score)</td>
<td>8</td>
<td>479</td>
<td>0.29</td>
<td>0.34 (0.05, 0.52)</td>
<td>-0.47</td>
<td>0.98</td>
<td>3</td>
<td>79</td>
<td>27 29 32 68 84 124 136 136</td>
</tr>
<tr>
<td>Corsi Blocks (visual span)</td>
<td>4</td>
<td>166</td>
<td>-0.29</td>
<td>0.34 (-0.62, 0.04)</td>
<td>-0.73</td>
<td>0.66</td>
<td>2</td>
<td>79</td>
<td>29 68 129 130</td>
</tr>
<tr>
<td>Trails B (time)</td>
<td>8</td>
<td>233</td>
<td>-0.28</td>
<td>0.39 (-0.56, -0.01)</td>
<td>-0.89</td>
<td>2.39</td>
<td>3</td>
<td>79</td>
<td>40 74 84 106 114 116 123 129</td>
</tr>
<tr>
<td>Stroop (colour naming/Stroop B)</td>
<td>3</td>
<td>92</td>
<td>0.22</td>
<td>0.35 (-0.62, 0.18)</td>
<td>-0.45</td>
<td>0.27</td>
<td>0</td>
<td>85</td>
<td>68 120 130</td>
</tr>
<tr>
<td>Stroop (interference)</td>
<td>3</td>
<td>70</td>
<td>0.18</td>
<td>0.36 (-0.33, 0.68)</td>
<td>-0.07</td>
<td>0.49</td>
<td>0</td>
<td>85</td>
<td>31 84 40 68</td>
</tr>
<tr>
<td>Stroop (errors—incorrect trial)</td>
<td>3</td>
<td>78</td>
<td>0.18</td>
<td>0.40 (-0.28, 0.63)</td>
<td>-0.13</td>
<td>1.09</td>
<td>0</td>
<td>85</td>
<td>28 30 35 72 74 78 86 91 110</td>
</tr>
<tr>
<td>Digit Span (forwards)</td>
<td>13</td>
<td>453</td>
<td>0.14</td>
<td>0.36 (-0.06, 0.63)</td>
<td>-0.80</td>
<td>2.73</td>
<td>0</td>
<td>92</td>
<td>111 119 123 129</td>
</tr>
<tr>
<td>Stroop (time—incorrect trial)</td>
<td>2</td>
<td>40</td>
<td>0.14</td>
<td>0.45 (-0.48, 0.77)</td>
<td>-0.23</td>
<td>0.37</td>
<td>0</td>
<td>92</td>
<td>27 130</td>
</tr>
<tr>
<td>Trails A (time)</td>
<td>5</td>
<td>162</td>
<td>-0.12</td>
<td>0.38 (-0.46, 0.21)</td>
<td>-0.73</td>
<td>0.32</td>
<td>0</td>
<td>92</td>
<td>40 74 84 111 129</td>
</tr>
<tr>
<td>Digit Span (backwards)</td>
<td>19</td>
<td>697</td>
<td>0.11</td>
<td>0.35 (-0.04, 0.27)</td>
<td>-1.16</td>
<td>1.08</td>
<td>0</td>
<td>92</td>
<td>20 28 30 40 47 72 74 77 80 88</td>
</tr>
</tbody>
</table>

ACE, Addenbrooke’s Cognitive Examination; Max d, maximum effect size; Mean $d_w$, mean weighted effect size; Min d, minimum effect size; N, number of studies contributing to the effect size; Nfs, fail safe N; %OL, per cent overlap between Alzheimer’s disease and frontotemporal dementia groups; SD $d_w$, SD of weighted effect size; 95% CI, 95% CI for means.

*Effect sizes weighted by the inverse variance.
dementia type. Every study was given a score according to the diagnostic criteria that were used to define each of their samples. As post-mortem neuropathological confirmation of AD or FTD provides the gold standard for a diagnosis of dementia, this was given a score of three (N_AD = 5, N_FTD = 5 studies), while studies using the National Institute of Neurologic, Communicative Disorders and Stroke-AD and Related Disorders Association (NINCDS-ADRDA) diagnostic criteria for AD57 and the Lund and Manchester criteria for FTD50 were given a score of two (N_AD = 82, N_FTD = 74 studies). Studies that used revised or similar criteria, particularly prior to the publication of these guidelines, were also given a score of two. Finally, a score of one was given to studies whose diagnostic criteria were less rigorous or not clearly specified (N_AD = 7, N_FTD = 15 studies). The diagnostic scores for studies therefore varied between 2 and 6.

The potential influence of three other methodological variables (participant age, years since diagnosis and education) was also examined. Mean age, years since diagnosis and years of education were calculated for each study (eg, M_AD-HTDAGE) for this purpose. This was done by combining the data from the AD and FTD groups for that study and weighting it by the sample sizes of the AD and FTD groups.

While there is considerable confusion within the literature regarding the best way to analyse the influence of methodological variables on effect sizes, Steel and Kammeyer-Mueller51 have demonstrated that a weighted least squares (WLS) multiple regression (using the inverse of the sampling error variance to weight the multiple regression) is less affected by the problems of heteroscedasticity and multicollinearity, which can affect the accuracy of these analyses. Heteroscedasticity occurs when the distribution of study sample sizes is skewed very large samples) and multicollinearity occurs when predictor variables are correlated with one another.51 A WLS multiple regression was therefore conducted, in addition to Pearson r correlations, to examine these methodological variables. Standard errors and confidence intervals for the results of the WLS multiple regression were adjusted according to the method outlined by Hedges.51

### RESULTS

#### Participants

The demographic characteristics for the participants in the 94 studies that were included in the current meta-analysis are shown in Table 2. Overall, 4684 participants were included in this study. Gender was only reported in 69 studies, providing data for 3480 cases (males: N_AD = 917, N_FTD = 737; females: N_AD = 1253, N_FTD = 573). When the demographic characteristics of the AD and FTD participants that were included in this meta-analysis were compared, the FTD participants were found to be significantly younger (t = 12.59, df = 86, p<0.001), more educated (t = -2.51, df = 60, p = 0.015) and had significantly higher MMSE scores (t = -5.41, df = 74, p<0.001) than the participants with AD. However, the two groups did not differ significantly in terms of the time since their diagnosis (t = 0.38, df = 36, p = 0.709) or Clinical Dementia Rating scores (t = 0.25, df = 18, p = 0.807).

Five studies had pathological confirmation of AD,56-58 82 used published criteria,54 25 26 31 40 42 44 59-129 and seven did not specify diagnostic criteria.100-106 For the diagnosis of FTD, five studies had pathological confirmation,54 25 26 31 40 and 74 used published criteria.51 25 26 31 40 42 44 59-66 78 109-114, 116 120-124 126-129 131 14 did not clearly specify the diagnostic criteria and one study used a combination of clinical observations and published criteria.19

There were not enough studies reporting data for premorbid intelligence (n = 6) or scores on the Dementia Rating Scale (n = 6) to report or analyse this information. Similarly, too few studies reported information about participants’ current medications (n = 19), much of which was not comparable, to meaningfully examine this information. Thirty-five studies reported information regarding the inclusion or exclusion of depressed participants. However, only six of these provided scores from depression rating scales.

#### Cognitive tests

The weighted effect sizes (d_w) for all measures (mean, SD, 95% CI, minimum, maximum), grouped according to test category (orientation and attention, memory, construction, verbal abilities and language, concept formation and reasoning, executive functioning and motor tasks, and general ability and intelligence) and rank ordered by size, are provided in Tables 3-10. Nfs and the percentage overlap between groups (%OL) are also provided, as are the number of studies (N) that used each measure, the number of participants that were assessed and the study references. The conclusions drawn from this study are based on the combined interpretation of these statistics. From a neuropsychological perspective, differential diagnoses are likely to be more accurate when there are large group differences (d) and there is limited overlap in the cognitive profiles (%OL). A clinician would also be more confident in their decision if the CI did not span zero (ie, groups differ significantly) and it is unlikely that the possibility of unpublished findings would draw a diagnosis into question (Nfs). Thus in order for a measure to be useful for differentiating between AD and FTD dementia, it had to meet the following criteria: (i) a large effect size (ie, d_w ≥ 0.8) and, consequently, a limited degree of group overlap (%OL <50), (ii) a confidence interval that did not span zero (CIs are affected by N and SD/range) and (iii) an Nfs score that was large enough to make it unlikely that there were that number of unpublished studies with non-significant findings in existence. As different tests were used with varying frequency, it was decided that the Nfs should be greater than the number of published studies that had used the test (ie, Nfs >Nstudies).

---

### Table 4: Perception: weighted Cohen's d effect sizes* for each test, rank ordered by size

<table>
<thead>
<tr>
<th>Test</th>
<th>N studies</th>
<th>N participants</th>
<th>Mean Cohen's d_w</th>
<th>SD d_w (95% CI)</th>
<th>Min d</th>
<th>Max d</th>
<th>Nfs</th>
<th>%OL</th>
<th>Max CI</th>
<th>Study references</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOSP position discrimination</td>
<td>2</td>
<td>57</td>
<td>-0.70</td>
<td>0.40 (-1.25, -0.14)</td>
<td>-1.34</td>
<td>0.00</td>
<td>5</td>
<td>57</td>
<td>28</td>
<td>109</td>
</tr>
<tr>
<td>VOSP dot counting</td>
<td>3</td>
<td>78</td>
<td>-0.68</td>
<td>0.54 (-1.29, -0.06)</td>
<td>-1.07</td>
<td>0.00</td>
<td>7</td>
<td>52</td>
<td>78</td>
<td>109</td>
</tr>
<tr>
<td>VOSP object decision</td>
<td>4</td>
<td>109</td>
<td>0.34</td>
<td>0.40 (-0.06, 0.73)</td>
<td>-0.75</td>
<td>0.99</td>
<td>3</td>
<td>79</td>
<td>78</td>
<td>110</td>
</tr>
<tr>
<td>VOSP cube analysis</td>
<td>4</td>
<td>98</td>
<td>-0.23</td>
<td>0.42 (-0.64, 0.17)</td>
<td>-0.61</td>
<td>0.52</td>
<td>1</td>
<td>85</td>
<td>78</td>
<td>110</td>
</tr>
<tr>
<td>VOSP incomplete letters</td>
<td>4</td>
<td>108</td>
<td>0.04</td>
<td>0.39 (-0.35, 0.43)</td>
<td>-0.83</td>
<td>0.47</td>
<td>0</td>
<td>100</td>
<td>78</td>
<td>110</td>
</tr>
</tbody>
</table>

Max d, maximum effect size; Mean d_w, mean weighted effect size; Min d, minimum effect size; N, number of studies contributing to the effect size; Nfs, fail safe N; SD d_w, SD of weighted effect size; %OL, per cent overlap between Alzheimer’s disease and frontotemporal dementia groups; VOSP, Visual Object and Space Perception Battery; 95% CI, 95% CI for means.

*Effect sizes weighted by the inverse variance.
Overall, the effect sizes ranged from a minimum of −0.05 for the Judgement of Line Orientation Test to a maximum of 1.39 for the Graded Naming Test. Thus there was considerable variation in the extent to which the AD and FTD groups differed on the measures used by the 94 studies that were included in the current meta-analysis. This is also reflected in the per cent overlap statistics, which indicate that there was 100% overlap between the AD and FTD groups on the least discriminating measure (Judgement of Line Orientation Test) and 32% overlap between the two groups for the most discriminating measure (Graded Naming Test). Also interesting is the fact that the cognitive tests reported here were used by between two (eg, Addenbrooke’s Cognitive Examination (ACE) subtests, Western Aphasia Battery (WAB) Comprehension) and 43 studies (Category Fluency). Thus while some measures have been used infrequently, others have been studied very extensively (eg, Category Fluency, Letter Fluency, Rey Complex Figure, Boston Naming Test). In terms of the statistical significance of the effect sizes, as indicated by CIs that do not span zero, there were 69 effect sizes that differed significantly from zero.

Of the 18 measures of orientation and attention, only the ACE orientation score met the study criteria (see table 3). This measure had a large and significant effect size, reflecting 41% overlap in the scores of the AD and FTD groups, with the AD group performing more poorly than the FTD group on this measure. The Nfs for this measure was also well in excess of the number of published studies using this measure, suggesting a high degree of confidence in this finding.

Only five measures of perception were used by two or more studies and these were all taken from the Visual Object and Space Perception battery (see table 4). None of these measures produced large effects, although position discrimination and dot counting revealed moderate group differences and CIs that did not span zero. However, these measures were associated with more than 50% overlap between the two groups and both had modest Nfs, indicating that the effect sizes could be reduced to a small effect size if there were 5−7 unpublished studies that had found small group differences. Therefore, none of the measures of perception accurately discriminated between persons with AD and FTD.

Memory was one of the most commonly assessed cognitive domains. In particular, the Rey Complex Figure Test and the Logical Memory substest from the Wechsler Memory Scale (WMS) were very commonly used (see table 5). The recognition and delayed recall scores of the Auditory Verbal Learning Test (AVLT), the delayed recall trial of the Rey Complex Figure task, the Recognition Memory Test (words), the retention and delayed recall scores of the Logical Memory substest from the WMS, the memory score from the ACE, the delayed recall score of the Visual Reproduction substest from the WMS and the total recall score from the Selective Reminding Test all had large and significant effect sizes and less than 50% overlap between groups. Persons with AD performed more poorly than those with FTD on all of these measures. The delayed recall of the Rey Complex Figure test, in particular, has been used widely (21 studies) and the Nfs suggest that it is unlikely that this result would be negated by non-significant unpublished findings.

The recognition score of the AVLT, the Recognition Memory Test for words, the retention score of the Logical Memory substest from the WMS and the memory score of the ACE have been less widely used and, while acceptable, their Nfs were substantially lower.

Regarding tests of verbal abilities and language, there were more tests that fell into this category than any other. Eight of these test scores yielded large effect sizes, 95% CIs that did not
span zero and large Nfs statistics, indicating that there were large and significant group differences on these measures and that these findings are unlikely to be negated by unpublished findings (see table 6). Specifically, these measures were: the Graded Naming Test, Word–Picture Matching, WAB Spontaneous Speech (fluency), the Pyramids and Palm Trees (word score, picture score), other picture naming tasks, WAB Comprehension and WAB Repetition. For all eight verbal measures, the FTD groups performed more poorly than the AD group. In addition to these eight measures, there were a number of other commonly used measures, such as the Pyramids and Palm Trees (score not specified) and letter fluency tasks, which showed moderate effect sizes, had 95% CIs that did not span zero and large Nfs statistics. However, because the effect sizes for these measures were relatively low (indicating small to medium differences in the AD and FTD means), there was an unacceptably high degree of overlap (53% and 67%, respectively) between the scores for the AD and FTD groups.

Of the seven measures of construction that were used by two or more studies, only the Beery Developmental Test of Visual Motor Integration demonstrated a large group difference, less than 50% overlap in scores, 95% CIs that did not span zero and a satisfactory Nfs (see table 7). As with the abovementioned memory tests, the effect size was negative, indicating that the AD groups performed more poorly on this task than the FTD groups.

Despite researchers using a number of different measures of concept formation and reasoning, none of them showed large group differences (ie, d > 0.8) (see table 8). Indeed, there was between 57% and 92% overlap in the scores of the AD and FTD groups on all 12 measures. Commonly used measures, such as the Wisconsin Card Sorting Test and Raven’s Progressive Matrices, did not successfully discriminate between AD and FTD when the findings of all studies that had used these measures were considered, although the minimum and maximum effect sizes for these tests indicate that there were individual studies that reported large effects for these tests.

When the classifications used by Lezak et al and Spreen and Strauss were used to group the cognitive tests, there were only four tests that fell into the category of executive function and motor performance (see table 9), although other tests that are
discriminating between AD and FTD. In other categories (e.g., verbal fluency tasks, Wisconsin Card Sorting Test, Wechsler Adult Intelligence Scale-Similarities),17 18 None of these met the criteria specified for discriminating between AD and FTD.

Finally, of the five measures of general ability and intelligence that were used by the studies included in this meta-analysis, the MMSE was the only measure that revealed a large negative effect (see Table 10). This indicates that AD patients were more impaired on this measure than FTD participants. Despite its widespread use, only those studies (n = 5) that had pathological confirmation of a patient’s diagnosis were used in the calculation of this effect size. This was done in order to avoid a situation where the MMSE was being used both in group allocation (independent variable) and as a measure of cognitive performance (dependent variable).

### Moderator variables

The influence of four methodological variables (age, years since diagnosis, education, diagnostic criteria) on the effect sizes was also examined using Pearson r correlation coefficients and a WLS multiple regression analysis. “Diagnostic criteria” refer to the rating given to the quality of the diagnostic criteria that was used to diagnose AD and FTD (i.e., pathological confirmation, published criteria or other criteria). The mean age, years since diagnosis, education and diagnostic criteria scores of those studies that reported this data were firstly correlated with the weighted mean effect sizes for these studies. Small and non-significant correlations were observed for age (r = −0.01, n = 90, p = 0.94), years since diagnosis (r = −0.10, n = 38, p = 0.57), education (r = 0.11, n = 63, p = 0.41) and diagnostic criteria (r = −0.08, n = 94, p = 0.46), indicating that these variables were not significantly related to the effect sizes. A WLS multiple regression was additionally performed on the data from the 22 studies that reported data for all four variables. The weighted mean effect size for a study was the dependent variable, the four moderator variables were the independent variables, and the inverse variance was used as the weighting variable. The results of this regression analysis are presented in Table 11. The model was non-significant (p = 0.98) and accounted for only 23% of the variance. Therefore, both the correlations and the WLS multiple regression indicate that age, years since diagnosis, education and diagnostic criteria did not significantly contribute to the effect sizes reported in this study. Thus there did not appear to be any systematic variation in the performance of the AD and FTD groups on the different tests of cognitive ability as a consequence of between-study differences in these variables, suggesting that it is acceptable to combine the results of studies that differed on these methodological variables.

### DISCUSSION

Overall, the data for this meta-analysis were obtained from 94 studies that examined the cognitive performance of 2936

---

**Table 7**  Construction: weighted Cohen’s d effect sizes* for each test, rank ordered by size

<table>
<thead>
<tr>
<th>Test Description</th>
<th>N studies</th>
<th>N participants</th>
<th>Mean Cohen’s d₀</th>
<th>SD d₀ (95% CI)</th>
<th>Min d</th>
<th>Max d</th>
<th>Nfs</th>
<th>%OL</th>
<th>Study references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beery Developmental Test of Visual Motor Integration</td>
<td>2</td>
<td>68</td>
<td>−0.95</td>
<td>0.39 (−1.49, −0.41)</td>
<td>−1.00</td>
<td>−0.86</td>
<td>8</td>
<td>45</td>
<td>31 64</td>
</tr>
<tr>
<td>ACE Visuospatial Clock drawing tests</td>
<td>2</td>
<td>170</td>
<td>−0.58</td>
<td>0.24 (−0.91, −0.25)</td>
<td>−0.65</td>
<td>−0.50</td>
<td>4</td>
<td>62</td>
<td>69 70 92 113</td>
</tr>
<tr>
<td>Rey Complex Figure (copy)</td>
<td>25</td>
<td>874</td>
<td>−0.43</td>
<td>0.36 (−0.57, −0.29)</td>
<td>−1.47</td>
<td>0.52</td>
<td>29</td>
<td>73</td>
<td>90 97 98 110 115 119 120 123 129 134 156</td>
</tr>
<tr>
<td>Other figure copying tasks</td>
<td>7</td>
<td>551</td>
<td>−0.37</td>
<td>0.24 (−0.55, −0.19)</td>
<td>−0.77</td>
<td>0.41</td>
<td>6</td>
<td>73</td>
<td>40 44 49 79 93 105 126</td>
</tr>
<tr>
<td>WAIS Block Design</td>
<td>6</td>
<td>443</td>
<td>−0.29</td>
<td>0.32 (−0.54, −0.03)</td>
<td>−0.84</td>
<td>0.29</td>
<td>3</td>
<td>79</td>
<td>62 74 86 94 156</td>
</tr>
<tr>
<td>WAIS Object Assembly</td>
<td>2</td>
<td>300</td>
<td>0.28</td>
<td>0.26 (−0.08, 0.63)</td>
<td>0.21</td>
<td>0.42</td>
<td>1</td>
<td>79</td>
<td>42 74</td>
</tr>
</tbody>
</table>

ACE, Addenbrooke’s Cognitive Examination; Max d, maximum effect size; Mean d₀, mean weighted effect size; Min d, minimum effect size; N, number of studies contributing to the effect size; Nfs, fail safe N; %OL, per cent overlap between the Alzheimer’s disease and frontotemporal dementia groups; SD d₀, SD of weighted effect size; WAIS, Wechsler Adult Intelligence Scale (all editions); WLS, weighted least squares.

---

**Table 8**  Concept formation and reasoning: weighted Cohen’s d effect sizes* for each test, rank ordered by size

<table>
<thead>
<tr>
<th>Test Description</th>
<th>N Studies</th>
<th>N Participants</th>
<th>Mean Cohen’s d₀</th>
<th>SD d₀ (95% CI)</th>
<th>Min d</th>
<th>Max d</th>
<th>Nfs</th>
<th>%OL</th>
<th>Study references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proverbs</td>
<td>2</td>
<td>111</td>
<td>0.66</td>
<td>0.56 (0.29, 1.05)</td>
<td>0.25</td>
<td>1.00</td>
<td>5</td>
<td>57</td>
<td>69 92</td>
</tr>
<tr>
<td>WAIS Comprehension</td>
<td>4</td>
<td>359</td>
<td>0.65</td>
<td>0.56 (0.35, 0.95)</td>
<td>0.40</td>
<td>0.80</td>
<td>9</td>
<td>57</td>
<td>60 100 129 130</td>
</tr>
<tr>
<td>Cognitive Estimation Test</td>
<td>2</td>
<td>94</td>
<td>−0.51</td>
<td>0.30 (−0.93, −0.09)</td>
<td>−0.67</td>
<td>−0.16</td>
<td>3</td>
<td>67</td>
<td>29 44 48 74 96 101 120 124 129 130</td>
</tr>
<tr>
<td>Raven’s Progressive Matrices</td>
<td>10</td>
<td>686</td>
<td>−0.47</td>
<td>0.28 (−0.64, −0.30)</td>
<td>−1.75</td>
<td>0.29</td>
<td>14</td>
<td>67</td>
<td>60 62 73 86 94 126</td>
</tr>
<tr>
<td>WAIS Similarities</td>
<td>6</td>
<td>435</td>
<td>0.37</td>
<td>0.31 (0.12, 0.62)</td>
<td>−0.08</td>
<td>0.75</td>
<td>5</td>
<td>73</td>
<td>31 110 123</td>
</tr>
<tr>
<td>WCST (errors)</td>
<td>3</td>
<td>109</td>
<td>−0.30</td>
<td>0.34 (−0.68, 0.08)</td>
<td>−0.56</td>
<td>0.20</td>
<td>1</td>
<td>79</td>
<td>63 94</td>
</tr>
<tr>
<td>WAIS Arithmetic Comprehension</td>
<td>2</td>
<td>306</td>
<td>0.29</td>
<td>0.25 (−0.06, 0.63)</td>
<td>−0.22</td>
<td>0.59</td>
<td>1</td>
<td>79</td>
<td>63 94</td>
</tr>
<tr>
<td>WAIS Picture Completion</td>
<td>2</td>
<td>308</td>
<td>0.25</td>
<td>0.25 (−0.10, 0.59)</td>
<td>0.08</td>
<td>0.53</td>
<td>0</td>
<td>79</td>
<td>20 31 48 74 76 78 86 88 101 110 115 121 131</td>
</tr>
<tr>
<td>WCST (perseverations)</td>
<td>12</td>
<td>326</td>
<td>0.19</td>
<td>0.41 (−0.05, 0.42)</td>
<td>−1.16</td>
<td>1.39</td>
<td>0</td>
<td>85</td>
<td>20 28 31 48 74 76 86 88 101 110 115 121 131</td>
</tr>
<tr>
<td>WCST (categories)</td>
<td>13</td>
<td>367</td>
<td>0.11</td>
<td>0.40 (−0.11, 0.33)</td>
<td>−1.41</td>
<td>1.83</td>
<td>0</td>
<td>92</td>
<td>63 94 102 123 129 131</td>
</tr>
<tr>
<td>WAIS Picture Arrangement</td>
<td>4</td>
<td>479</td>
<td>−0.07</td>
<td>0.26 (−0.32, 0.19)</td>
<td>−0.44</td>
<td>0.25</td>
<td>0</td>
<td>92</td>
<td>44 62 86 94</td>
</tr>
</tbody>
</table>

Max d, maximum effect size; Mean d₀, mean weighted effect size; Min d, minimum effect size; N, number of studies contributing to the effect size; Nfs, fail safe N; %OL, per cent overlap between the Alzheimer’s disease and frontotemporal dementia groups; SD d₀, SD of weighted effect size; WAIS, Wechsler Adult Intelligence Scale (all editions); WCST, Wisconsin Card Sorting Test; 95% CI, 95% CI for means.

*Effect sizes weighted by the inverse variance.
persons with AD and 1748 persons with FTD. Forty-seven cognitive tests, which yielded 115 different scores, were used by more than one study. While the participants with AD were, on average, older and less educated than the FTD participants, the two groups were comparable in terms of time since diagnosis. The lower age of the FTD group is consistent with the fact that FTD tends to have an earlier onset. FT participants also had higher MMSE scores than the AD participants. However, this difference in MMSE scores is likely to reflect the fact that this measure was originally designed to detect the deficits associated with AD (e.g., memory, praxis) rather than FTD.

The current meta-analysis found that a wide variety of tests have been used in research that has examined the cognitive deficits associated with AD and FTD and that there is considerable variation in the ability of these measures to distinguish between AD and FTD. In order for a cognitive test to be useful for differential diagnosis, it was argued that it must be able to distinguish between the performance of these two groups (indicated by large effects and small %OL). There should also be a high degree of confidence in the accuracy with which the measure distinguishes between the two groups (measured by the 95% CI) and we need to be assured that the conclusions drawn from the research literature are not systematically biased by the tendency to publish statistically significant findings (Nfs statistic). When these three criteria were applied to all of the measures analysed in this study, there were only 19 out of 115 measures that met these criteria. More specifically, persons with AD performed more poorly than those with FTD on 12 of the 19 measures, all of which assessed orientation, memory or general cognitive ability. These tests were (in order of discriminative ability): the ACE orientation subtest, the MMSE, the Beery Developmental Test of Visual Motor Integration and nine measures of memory (AVLT recognition and delayed recall scores, Rey Complex Figure delayed recall score, Recognition Memory Test (words), WMS Logical Memory per cent retention and delayed recall scores, ACE memory subtest, WMS Visual Reproduction delayed recall score and Selective Reminding Test total recall).

In contrast, persons with FTD performed more poorly than those with AD on seven measures of verbal ability (Graded Naming Test, Word–Picture Matching, WAB Spontaneous Speech (fluency), Pyramids and Palm Trees (word score, picture score), other picture naming tasks and WAB Comprehension). Importantly, however, there was still between 32% and 48% overlap in the scores of the AD and FTD groups on all 19 of these measures. This level of overlap suggests that, even when the most discriminating cognitive tests are used, the differential diagnosis of AD and FTD remains problematic. Failure to find tests that clearly discriminate between AD and FTD confirms previous research which found that the performance of AD and FTD participants did not differ significantly on a large range of tests.27

The finding that there were nine measures of memory that were among the most useful for differentiating between AD and FTD is consistent with research and clinical observations that memory is differentially affected in AD and FTD. Numerous previous studies have reported that AD patients experience greater memory problems than those with FTD. In contrast, persons diagnosed with FTD had more difficulty with tests of verbal ability and language than those with AD, as demonstrated by eight large effect sizes in this domain. This supports previous research which indicated that FTD is associating with language and small %OL. There should also be a high degree of confidence in the accuracy with which the measure distinguishes between the two groups (measured by the 95% CI) and we need to be assured that the conclusions drawn from the research literature are not systematically biased by the tendency to publish statistically significant findings (Nfs statistic). When these three criteria were applied to all of the measures analysed in this study, there were only 19 out of 115 measures that met these criteria. More specifically, persons with AD performed more poorly than those with FTD on 12 of the 19 measures, all of which assessed orientation, memory or general cognitive ability. These tests were (in order of discriminative ability): the ACE orientation subtest, the MMSE, the Beery Developmental Test of Visual Motor Integration and nine measures of memory (AVLT recognition and delayed recall scores, Rey Complex Figure delayed recall score, Recognition Memory Test (words), WMS Logical Memory per cent retention and delayed recall scores, ACE memory subtest, WMS Visual Reproduction delayed recall score and Selective Reminding Test total recall).

In contrast, persons with FTD performed more poorly than those with AD on seven measures of verbal ability (Graded Naming Test, Word–Picture Matching, WAB Spontaneous Speech (fluency), Pyramids and Palm Trees (word score, picture score), other picture naming tasks and WAB Comprehension). Importantly, however, there was still between 32% and 48% overlap in the scores of the AD and FTD groups on all 19 of these measures. This level of overlap suggests that, even when the most discriminating cognitive tests are used, the differential diagnosis of AD and FTD remains problematic. Failure to find tests that clearly discriminate between AD and FTD confirms previous research which found that the performance of AD and FTD participants did not differ significantly on a large range of tests.27

The finding that there were nine measures of memory that were among the most useful for differentiating between AD and FTD is consistent with research and clinical observations that memory is differentially affected in AD and FTD. Numerous previous studies have reported that AD patients experience greater memory problems than those with FTD. In contrast, persons diagnosed with FTD had more difficulty with tests of verbal ability and language than those with AD, as demonstrated by eight large effect sizes in this domain. This supports previous research which indicated that FTD is associating with language impairment. In addition, numerous previous studies have reported that persons with FTD have more difficulty with executive tasks than AD patients. While executive tasks would be expected to discriminate between AD and FTD, given the frontal pathology associated with FTD, executive tasks were expected to discriminate between AD and FTD, given the frontal pathology associated with FTD. However, the frontal assessment battery (Max d, maximum effect size; Mean d_a, mean weighted effect size; Min d, minimum effect size; N, number of studies contributing to the effect size; Nfs, fail safe N; %OL, per cent overlap between the Alzheimer’s disease and frontotemporal dementia groups; SD d_a, SD of weighted effect size; 95% CI, 95% CI for means).

*Effect sizes weighted by the inverse variance.

### Table 9: Executive and motor functions: weighted Cohen’s d effect sizes* for each test, rank ordered by size

<table>
<thead>
<tr>
<th>Test</th>
<th>N Studies</th>
<th>N Participants</th>
<th>Mean Cohen’s d_a</th>
<th>SD d_a (95% CI)</th>
<th>Min d</th>
<th>Max d</th>
<th>Nfs</th>
<th>%OL</th>
<th>Study references</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAB</td>
<td>4</td>
<td>314</td>
<td>0.43</td>
<td>0.25 (0.19, 0.67)</td>
<td>-0.12</td>
<td>1.29</td>
<td>5</td>
<td>73</td>
<td>69 72 76 79 84</td>
</tr>
<tr>
<td>FAB (go no go)</td>
<td>2</td>
<td>143</td>
<td>0.29</td>
<td>0.26 (-0.07, 0.64)</td>
<td>0.25</td>
<td>0.33</td>
<td>1</td>
<td>79</td>
<td>69 72 76 79 84</td>
</tr>
<tr>
<td>Design Fluency</td>
<td>7</td>
<td>306</td>
<td>-0.06</td>
<td>0.32 (-0.30, 0.17)</td>
<td>-0.51</td>
<td>0.61</td>
<td>0</td>
<td>92</td>
<td>126 139</td>
</tr>
<tr>
<td>Frontal Composite Error Score</td>
<td>2</td>
<td>170</td>
<td>-0.13</td>
<td>0.24 (-0.46, 0.20)</td>
<td>-0.50</td>
<td>0.56</td>
<td>0</td>
<td>92</td>
<td>126 139</td>
</tr>
</tbody>
</table>

### Table 10: General ability and intelligence: weighted Cohen’s d effect sizes* for each test, rank ordered by size

<table>
<thead>
<tr>
<th>Test</th>
<th>N Studies</th>
<th>N Participants</th>
<th>Mean Cohen’s d_a</th>
<th>SD d_a (95% CI)</th>
<th>Min d</th>
<th>Max d</th>
<th>Nfs</th>
<th>%OL</th>
<th>Study references</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMSE</td>
<td>5</td>
<td>168</td>
<td>-0.98</td>
<td>0.37 (-1.31, -0.65)</td>
<td>-1.29</td>
<td>-0.01</td>
<td>19</td>
<td>45</td>
<td>54 58 68 72 98 130</td>
</tr>
<tr>
<td>WAIS Verbal IQ</td>
<td>5</td>
<td>427</td>
<td>0.51</td>
<td>0.29 (0.26, 0.77)</td>
<td>0.14</td>
<td>0.92</td>
<td>8</td>
<td>67</td>
<td>58 68 72 98 130</td>
</tr>
<tr>
<td>WAIS Full Scale IQ</td>
<td>6</td>
<td>433</td>
<td>0.30</td>
<td>0.31 (0.05, 0.56)</td>
<td>-0.23</td>
<td>0.62</td>
<td>3</td>
<td>79</td>
<td>58 68 72 98 130</td>
</tr>
<tr>
<td>ACE</td>
<td>4</td>
<td>171</td>
<td>-0.18</td>
<td>0.34 (-0.51, 0.14)</td>
<td>-0.70</td>
<td>0.57</td>
<td>0</td>
<td>85</td>
<td>58 68 72 98 130</td>
</tr>
<tr>
<td>WAIS Performance IQ</td>
<td>5</td>
<td>428</td>
<td>0.16</td>
<td>0.29 (-0.09, 0.41)</td>
<td>-0.56</td>
<td>0.69</td>
<td>0</td>
<td>85</td>
<td>58 68 72 98 130</td>
</tr>
</tbody>
</table>

ACE, Addenbrooke’s Cognitive Examination; Max d, maximum effect size; Mean d_a, mean weighted effect size; Min d, minimum effect size; MMSE, Mini-Mental State Examination; N, number of studies contributing to the effect size; Nfs, fail safe N; %OL, per cent overlap between the Alzheimer’s disease and frontotemporal dementia groups; SD d_a, SD of weighted effect size; WAIS, Wechsler Adult Intelligence Scale (all editions); 95% CI, 95% CI for means.

*Effect sizes weighted by the inverse variance.
participants, years since diagnosis, years of education and the
diagnostic criteria that were used were not related to the size of
the effects reported by these studies. The finding that years
since diagnosis was not related to the group differences is
surprising given that the symptoms associated with both AD
and FTD vary with the stage of illness, with cognitive
performance declining over time. However, demographic
variables such as age, gender, years since diagnosis and
education were not consistently reported. For example, only
approximately 40% of the studies reported years since
diagnosis. Thus there is the potential for significant but
unreported variation in these variables. Also important is the
fact that mean effect sizes were calculated for each study in
order to complete these analyses. Thus the effect sizes for all of
the individual measures that were used by a given study were
averaged. If a study yielded both small and large effect sizes for
different tests, these were combined to calculate a mean effect
size, which was then correlated with the moderator variables. A
more useful analysis would involve analysing each measure
separately by correlating the effect sizes from all of the studies
that used a particular measure with the age, education and
years since diagnosis data for those studies. However, this was
not possible as the maximum number of studies that used any
given measure was 43. When those studies that did not provide
data for the moderator variables were excluded, the sample size
was reduced to an unacceptably low number (ie, 22). Thus the
current moderator analysis was necessarily limited. Similarly,
current medications and whether depressed participants were
excluded was rarely reported. Therefore, these variables could
not be examined, despite the potential for medications and
depression to affect cognitive performance. It is possible that
these factors contributed to the effect sizes. It is, therefore,
essential that this information is included in future publica-
tions of primary research to allow an accurate and detailed
synthesis of the research findings.

There are a number of limitations to the current meta-
thesis that warrant consideration. Firstly, rating inventories
were not included in the current meta-analysis as they are not
classified as objective cognitive tests. However, given the
behavioural nature of FTD, a meta-analysis of these beha-
avioural rating scales may yield useful findings. In addition,
assessments of social cognition, personality and empathy
measures, which have been recommended in the differential
diagnosis of dementia, did not fall within the scope of
the current meta-analysis but may also be worthy of
consideration.

Secondly, it is important to note that the current meta-
analyses specifically examined AD and FTD. The overlapping
cognitive features of these types of dementia are highlighted by
the fact that few tests clearly discriminated between these two
groups. While greater differences would be expected if the
cognitive performance of the AD and FTD groups were
compared with healthy controls, such an analysis would not
adequately address the question of differential diagnosis.

Thirdly, FTD is described using a variety of terms, which were
combined in the current meta-analysis. This may have
increased the variability in the test performance of the FTD
group. Although it would be desirable to analyse these
subgroups separately, the labels and the diagnostic criteria
that are applied to these subgroups are not used consistently,
thereby precluding a more specific analysis of potential
subtypes. Moreover, more research is needed on each of these
subgroups if they are to be analysed separately.

While this analysis indicates that many test scores do not
clearly discriminate between AD and FTD, there is some
evidence that qualitative aspects of performance may better
distinguish the two groups. For example, a recent study
indicated that the consideration of performance characteristics
and specific error types increased the accuracy of the
differential diagnosis of AD and FTD. However, given that
these aspects of performance are often not quantified,
discrimination between AD and FTD is largely reliant on
subjective clinical observations.

Finally, the effect sizes derived from measures used by single
studies were not included in the current meta-analysis.
Technically, a meta-analysis requires two or more studies in
order to aggregate primary research findings. One conse-
quence of this is that there may be interesting and innovative
measures that discriminate between groups but are not in
widespread use and were not included in the current analysis.

CONCLUSIONS
In summary, the findings of this meta-analysis suggest that the
neuropsychological tests that best discriminate between AD
and FTD are ACE orientation, ACE memory, recognition and
delayed recall of the AVLT, delayed recall of the Rey Complex
Figure, the words version of the Recognition Memory Test,
WMS Logical Memory (per cent retention and delayed recall),
WMS Visual Reproduction Test (delayed), total recall of the
Selective Reminding Test, Graded Naming Test, Word–Picture
Matching, WAB Spontaneous Speech (fluency), Pyramids and
Palm Trees (word score, picture score), picture naming tasks,
WAB Comprehension, the Beery Developmental Test of Visual
Motor Integration and MMSE. However, it is important to note
that none of the tests showed acceptably low overlap between
the scores of the two groups to confidently make a differential
diagnosis. Therefore, these cognitive tests must be used
cautiously and in conjunction with other diagnostic informa-
tion, such as medical history, behavioural observations,
imaging and information from relatives when making a
differential diagnosis.

ACKNOWLEDGEMENTS
We would like to acknowledge the statistical expertise of Mr B Willson
and Dr M Hutchinson in this research. The preliminary findings from
this study were presented at the International Neuropsychological

Authors’ affiliations
A D Hutchinson, J L Mathias, School of Psychology, University of Adelaide,
Adelaide, Australia

Competing interests: None.

REFERENCES
1. Prigerson HG. Costs to society of family caregiving for patients with end-stage
2. Davidoff DA. The recognition and treatment of dementia. In: Ellison JM,
Weinstein C, Hodel-Malinowsky T, eds. The psychotherapist’s guide to
neuropsychiatry: Diagnostic and treatment issues. Washington, DC: American
926 Hutchinson, Mathias


